共主题网络方法及应用*

钮　亮
(中国计量大学经济与管理学院　杭州　310018)

摘要:【目的】通过构建共主题网络，对主题之间的关系进行分析，优化主题包含的词项。【方法】将"文档–主题"二分图依照加权投影规则生成共主题网络，使用介数中心性和主题概率结合的方法测度共主题网络中重点主题，通过 GN 算法对主题网络进行社区分割，使用相关度方法优化主题词项。【结果】将共主题网络与基于 JSD 的 K-means 方法进行比较发现，两者在三种主题数(最优主题数 28 和随机主观主题数 20，30)测试下产生的聚类数目都相同，聚类内容的一致程度分别达到 100%、95%、87%。【局限】其他社区分割方法共主题网络未能全面涉及。【结论】共主题网络照顾到了高维数据的需要，能够探查出文档中哪些主题是重要主题，哪些主题联系紧密。

关键词: 共主题网络　LDA　社区分割　K-means
分类号: G250

1　引　言

科技文献资源的利用一直以来都受到学术界的重视，以往的研究一般是利用共词分析方法对科技文献进行计量分析，集中在分析对象的改进、指标改进、可视化方法调整等方面[1]。但共词分析方法由于难以发现文档中潜在的语义联系，无法满足用户对科技信息深层次的需求。在自然语言处理领域提出了 LDA 主题模型[2]，由于其围绕语义问题进行词项分配，很快被引入到科技文献的计量分析之中[3-4]。并在其基础上形成一些比较经典的扩展，如 AT 模型[5]、TOT 模型[6]、CTM 模型[7]等。

尽管 LDA 在科技文献挖掘方面取得了一定的成绩，但是传统 LDA 模型仍然存在两个明显的问题:

(1) LDA 模型训练语料后形成的主题之间缺乏联系。传统 LDA 模型在解释文档时常常选择一个概率分布最高的主题来说明文档，然而一个文档有时候不仅仅只体现一个主题内容，它可能由若干个主题构成，因此传统 LDA 模型对于这些共同出现在文档中的主题的关系如何，在表达文档意义时哪些主题是更重要的，并未给以解释。为了表达主题之间的关系，也有一些文献引入了主成分分析方法和聚类方法，将多个维度的主题压缩成两维来计算，并通过多维标度来展示[8-9]。其中聚类的距离测度采用 KL 距离(Kullback-Leibler Divergence)[5,10]、JSD(Jensen-Shannon Divergence)[11-13]、余弦相似度[14]方法来实现。这些文献存在的问题是主成分分析方法将高维数据压缩成两维数据来处理，忽略了高维数据的复杂性。聚类中距离测度到底要选择哪种方法存在困扰。复杂网络很好地照顾到了高维数据的需要，也能够利用社区分割方法完成数据的聚类问题。因此在探测主题关系时，根据主题在文档中的共现情况，构建文档–主题二分图网络，并通过投影形成主题网络，实现对主题关系的探测。主题模型和复杂网络结合的文献较少，文献[10]和文献[15]把合作网络中的每个用户看作文档，每个用户的所有合作者看作该文档的词项，使用主题模型的方法处理合作网络用户的聚类问题，这些网络数据属于LDA模型的语料准备。文献[16]将复杂网络社区分割的模块度作为隐

通讯作者: 钮亮，ORCID: 0000-0001-6934-4416，E-mail: niutyut@126.com。

*本文系国家自然科学基金项目"碳排放规则下供应链成员企业行为及网络均衡协调研究"(项目编号: 71402173)、浙江省高校人文社会科学重点研究基地"决策科学与创新管理"项目"物流配送 VRP 模型、算法及其在 GIS 中的应用研究"(项目编号: RWSKZD03-201207)和浙江省产业发展政策研究中心、浙江省标准化与知识产权管理研究基地项目"FDI 视角下浙江省物流产业竞争力的提升策略研究"(项目编号: SIPM3222)的研究成果之一。

形变量以增强 LDA 模型的性能，与 AT 模型、TOT 模型的性质类似，对生成的主题关系并未进一步讨论。

(2) 传统LDA模型的主题词项构成中经常会存在一些不重要的，甚至是不太有关联的词项。模型输出结果需要领域专家来确定和修改哪些词项是有意义的[17-18]，无法自动完成词项的选择[19-20]，后来有一些模型通过引入一些外在的隐性变量来增强构成主题的词项的精度，例如 AT 模型引入作者这个隐形变量以增强 LDA 模型的精度。TOT 模型引入时间因素，将时间看作连续的可观测变量来增强词项精度。但由于这些扩展模型与LDA的生成机制是一样的，产生的主题词项中依旧无法避免存在不太关联的词项。

针对传统 LDA 存在的问题，本文将完成两项内容：

(1) 通过文档–主题二分图投影构建共主题网络，通过社区分割和中心性测度探测主题关系和重要主题；

(2) 优化主题词项选择，并遴选出与主题相关的词项。

2 理论与方法

2.1 LDA 主题模型

LDA 主题模型是一种非监督的机器学习方法，采用词袋(Bag of Words)表示的方法，这种方法将每一篇文档视为一个词频向量，从而将文本信息转化为易于建模的数字信息。LDA 的建模方式认为每一篇文档代表了一些主题所构成的一个概率分布，而每一个主题又代表了很多词项所构成的一个概率分布。如果有 T 个主题，则给定文档中 w_i 词项的概率如下：

$$P(w_i) = \sum_{j=1}^{T} P(w_i|z_i = j)P(z_i = j) \qquad (1)$$

其中，z_i 是潜在变量，表明第 i 个词汇 w_i 是从该主题得出的。$P(w_i|z_i = j)$ 是词汇 w_i 属于主题 j 的概率，$P(z_i = j)$ 给出主题 j 属于当前文本的概率。直观上，$P(w|z)$ 揭示的是哪些词对一个主题是重要的，而 $P(z)$ 是在一个文档中的主题分布。主题的内容反映在 $P(w|z)$ 中，一篇文章的构成依赖于主题分布 $P(z)$。例如一个期刊在“统计学习”和“机器学习”栏目中发布了若干论文，那么就认为词项的概率分布是围绕“统计学习”和“机器学习”的，它的主题内容反映在 $P(w|z)$ 中。在“统计学习”主题中，“方差”、“偏差”、“回归”这些词项就有很高的概率，而“统计学习”和“机器学习”主题就反映在 $P(z)$ 中。事实上，这种文章由若干主题构成，若干主题又由词项构成的情况是一个标准的贝叶斯分类问题。给定 D 个文档，这些文档包含 T(通过反复试验等方法事先给定)个主题且这些主题又由词汇表中 W 个独立的词项构成。其中，$P(w|z)$ 表示每个主题与词汇表中的 W 个词项的一个多项分布相对应，将这个多项分布记为 Φ，即 $P(w|z = j) = \Phi_w^{(j)}$。语料库中的 D 篇文档与 T 个主题的一个多项分布相对应，将该多项分布记为 θ，对于给定文档 D 中的某个 d 来说，$P(z = j) = \theta_j^{(d)}$。给定词项 $w=\{w_1, w_2, \cdots, w_n\}$，其中每一组 w_i 属于特定文档 d_i，对文档 d 中的每一个词项，从该文档所对应的多项分布 θ 中抽取一个主题 z，再从主题 z 所对应的多项分布 Φ 中抽取一个词项 w，将这个过程重复 N_d 次，就产生了文档 d，这里的 N_d 是文档 d 的词项总数。为了求解出主题，两个参数需要推断：“文档–主题”分布 θ、“主题–词项”分布 Φ。推断方法主要有 EM 算法和 Gibbs 抽样法。

2.2 共主题网络构建

(1) 文档–主题二分图及其投影

大多数网络是由一种节点类型组成的单模式网络，事实上还存在一种二分图网络，这种网络的节点属于不同的节点集，边是由这些不同类型的节点集合中的节点连接在一起的。针对文档–主题二分图网络，其中一种节点类别是文档，一种节点类别是主题。对文档–主题二分图网络进行形式化：设 G=<V, E> 且 $X \cup Y = V, X \cap Y = \phi$，使得 G 的每条边的两个端点一个属于 X，一个属于 Y，记为<X, Y, E>。其中 X 代表文档，Y 代表 X 文档选取的那些主题，选择规则是选取文档 X 中大于平均概率的主题作为代表文档的主题来构建文档–主题二分图网络。

二分图网络如果不做转化很少能够被分析，是由于大多数网络的测度手段为单模式图设计，只有很少的一些设计是解决二分图的，因此需要将二分图投影为单模式图。节点集 X 的投影规则为：如果节点集 X 中的任意两个节点与节点集 Y 中的某个节点都相连，那么就让节点集 X 的这两个节点连边。节点集 Y 的投影规则反之亦然。二分图及其投影如图 1 所示，其中

二分图为(a), X 节点投影为(b), Y 节点投影为(c)。

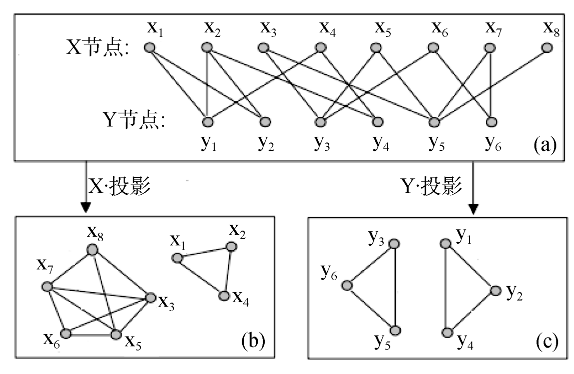


图 1 二分图及其投影

单模式图投影方法非常实用并被广泛使用，但是它的构造方式丢失了很多原始二分网络结构包含的信息。在映射中只是将同类的两个顶点做了连接，却没有考虑这两个顶点到底属于多少个群组。通过给投影赋予权重，可以在投影中保留这类信息。通常是将投影网络中的两个顶点间边的权重设置为它们共同属于的另一个群组的数目[21]; Newman 在科学家–论文二分图转化为科学家单模式图中认为，群组数目忽略了作者的贡献程度，作者的权重应该随着一篇论文合著作者数量的不同而有所不同[22]; Zhou 等认为 Newman 的方法[22]忽略了独著作者在投影中的重要性，他从资源分配的影响出发设计二分图投影后的权重[23] 。

本文的投影规则是将主题作为一个节点，两个主题如果出现在同一个文档之中，则在这两个主题之间建立一个连边，说明这两个主题存在相关关系，如果一个文档有 n 个主题，那么就产生 n(n–1)/2 种两两相关关系。当两个特定的主题 T1 和 T2 同时出现在多个文档中时，对 T1 和 T2 不再重复连接，但这样二分图中的文档节点就被忽略了。为了在投影中反映二分图的文档节点，同时又反映主题之间联系的紧密程度，设定主题 T1 和 T2 之间的权重如下:

$$W_{T1T2}=\sum_{k=1}^{g}\frac{\delta_{T1}^{k}\delta_{T2}^{k}}{n_k-1} \tag{2}$$

其中，g 表示文档数目，当主题 T1 在文档 k 中出现时，δ_{T1}^{k} 等于 1，否则为 0; n_k 表示文档 k 的主题数目。图 2 是该定义的一个示例，主题 T1 和 T2 在三个文档中都出现过，其中第一个文档有 4 个主题，第二文档有 2 个主题，第三个文档有 3 个主题，于是 T1 和 T2 在三个文档中的关系强度分别是 1/3、1、1/2，所以 T1 和 T2 总的关系强度是 1/3 + 1 + 1/2 = 11/6，这一关系强度就是主题 T1 和 T2 连边的权重。依照这种投影计算规则，对主题进行单模式图投影，生成加权的共主题网络。

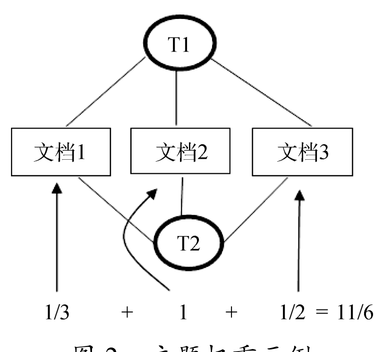


图 2 主题权重示例

(2) 节点重要性计算

共主题网络节点重要性探测由两方面构成：基于拓扑结构的共主题网络的节点中心性；基于LDA模型训练的基于词项分布的主题概率大小。节点中心性的探测有度中心性、介数中心性、接近中心性、特征向量中心性、k-壳与 k-核[24]。本文采用介数中心性和主题自身的词项概率分布来探测主题的重要程度。之所以将两种相结合是因为介数中心性解决的是任何一个节点达到其他节点的最短路径必然要经过的节点。介数中心节点起到建立其他节点彼此关系的作用，显然地位是非常重要的。但是仅仅有介数中心性，就无法体现其他节点自身的重要性，因为网络中很多节点本身并不担当连接其他节点桥梁的作用，因此节点自身的非拓扑性质的重要性需要体现。针对主题网络而言，这个非拓扑属性就是构成主题内容的词项的概率分布。因此在计算共主题网络中主题节点的重要程度时，本文将介数中心性和词项的概率分布结合在一起。节点 V_i 的介数中心性是网络中所有最短路径中经过该点的数量，公式如下:

$$B(i)=\sum_{u\neq w\neq i}\frac{\sigma_{uw}(i)}{\sigma_{uw}} \tag{3}$$

其中，σ_{uw} 是节点 V_u 和 V_w 之间的最短路径数量，$\sigma_{uw}(i)$ 是经过节点 V_i 的 V_u 和 V_w 之间的最短路径数量。介数越高，则节点越处于中心地位。当一个节点不在任何一条最短路径上时，其中心性为 0。

节点 V_i 的主题词项概率分布公式如下:

$$P(i) = \sum_{d=1}^{D} \theta_{di} \quad (4)$$

其中, θ_{di} 为文档 d 的话题主题节点 V_i 的多项式分布, D 为文档集合。

最终共主题网络中的主题节点强度由介数中心性、主题节点概率分布共同构成。由于数据之间量纲的不相同, 采用 Min-Max 标准化将其转化为无量纲的纯数值, 公式如下:

$$V_i = c\left(\frac{B(i) - \min(B(i))}{\max(B(i)) - \min(B(i))} + \frac{P(i) - \min(P(i))}{\max(P(i)) - \min(P(i))}\right) \quad (5)$$

其中, c 为调节系数, 调节主题节点的可视化显示大小。主题网络中主题节点的面积大小取决于 V_i 的值。从公式(5)可以看出 V_i 越大, 主题重要程度越高, 在主题网络中主题节点显示的面积也就越大。

(3) 共主题网络聚类

尽管共主题网络通过边权关系揭示了主题节点之间的关系, 突出了哪个主题是重要的, 但它们揭示的是两两主题之间的关系。而多个主题是否同属一种类别需要通过聚类进行揭示。针对共主题网络, 本文采取社区分割技术。对社区结构进行划分常用的方法有两类: 图论算法(包括谱平分法、随机游走算法、派系过滤法等)和层次聚类算法(凝聚算法和分裂算法)。前者的代表为基于贪婪算法思想的凝聚算法, 也称CNM算法[25]; 后者的代表为基于边介数的 GN 算法[26]。CNM 算法适用于大规模网络的社区分割, GN 算法只适用于中小型规模的网络。由于共主题网络节点边数少, 所以采用 GN 算法实现社区分割。GN 算法是一种分裂型的社区结构发现算法。该算法根据网络中社区内部高内聚、社区之间低内聚的特点, 逐步去除社区之间的边, 取得相对内聚的社区结构。算法用边介数的概念探测边的位置, 计算所有边的介数中心度, 最高介数的边被移除, 反复迭代计算剩下的边介数直到边介数低于某个阈值 μ, 算法停止。伪代码如下:

```
Input: A weighted or unweighted graph G = (V, E) , Threshold μ
Output: A list of clusters
while |E(G)| >0 do
  C_{u,v} - betweenness centrality of edge (u, v)
  Calculate C_{u,v} for all (u, v) ∈ E(G)
  maxBetweennessEdge = (x, y) : C_{x,y} is minimum over all (x, y)
  in E(G)
  maxBetweennessValue = C_{x,y}
  if maxBetweennessValue ≥ μ then
    E(G) = E(G) − {maxBetweennessEdge}
  else
    Break out of loop
  end
end
return Connected components of modified G
```

由于 GN 算法是反复迭代寻找边介数最大的值并移除, 因此无法判断算法终止位置, 而且还会重复计算节点的最短路径, 时间复杂度高。理论上无法自动确定最后会分割为多少个社区。社区的确定需要对阈值 μ 进行调试。因此需要有一种度量的方法, 判断不同阈值 μ 下面产生的结果是不是最佳的结果。为此 Newman 引入了模块度 Q 的概念来评价社区结构划分的质量[27-28]。但模块度只能用来衡量一个社区的划分是不是相对比较好, 之所以说相对是因为准确最优的模块度优化算法在计算上是困难的[29]。因此, 在计算模块度 Q 值时, Q 取值最大的时候就认为是网络较理想的划分。Q 值的范围在 0-1 之间, Q 值越大说明网络划分的社区结构准确度越高。模块度 Q 的计算公式如下:

$$Q = \frac{1}{2m}\sum_{i,j}\left[A_{ij} - \frac{k_i k_j}{2m}\right]\delta(c_i, c_j) \quad (6)$$

其中, A_{ij} 是 i 行 j 列邻接矩阵 A 的元素。k_i 是连接到节点 i 的度, k_j 是连接到节点 j 的度。c_i, c_j 表示节点 i 和 j 所在的两个社区。如果 i 和 j 在同一个社区, 则函数 $\delta(c_i, c_j) = 1$, 否则 $\delta(c_i, c_j) = 0$, $m = \frac{1}{2}\sum_{ij} A_{ij}$ 为网络中边的总数。GN 算法直接对模块度 Q 值进行最优化以寻找最佳社区结构, 它以使 Q 值增大最快或减少最慢为目标将社区一步一步地融合。在这个过程中会出现 Q 的峰值 Q_{max}, 其中 Q_{max} 对应的社区结构就是最佳社区结构。

2.3 主题词项优化

LDA 主题模型主要是对两个参数进行推断: "文档–主题"分布 θ 和"主题–词项"分布 Φ。共主题网络解决的是 θ 的关系问题, 由于共主题网络中的节点主题词项依赖于Φ, 因此Φ的优化直接决定了节点主题的被解释情况。传统主题模型训练后的主题里经常会存在一些不重要的, 甚至是不太相关的词项。因此模型

输出结果需要领域专家来确定和修改那些词项是有意义的。当前主题质量评价完全依赖专家对给定主题中词项的甄别。为了提高自动选择主题词项的能力，排除构成主题中的某些不相干的词项，文献[30]使用一种称为提升度法(Lift)的内在度量方法来排序主题中的词项，该提升度法被定义为某个主题中的词概率占据该词项在整个语料中边际概率的比率。文献[31]根据词项在主题中出现的频率和独占性进行排序。文献[9]结合提升度和独占性方法提出一种相关度方法(Relevance)，其中某个词语主题的相关性由 λ 参数来调节。如果 λ 接近 1，那么在该主题下频繁出现的词和主题更相关，正是文献[30]所讨论的；如果 λ 越接近 0，那么该主题下特殊、独有(Exclusive)的词和主题更相关，正是文献[31]所讨论的。考虑到文献[9]在主题词项产生中的精度，因此构建主题词项网络节点所需要的词项由主题词项相关性构建，公式如下：

$$r(w,k|\lambda)=\lambda\log(\phi_{kw})+(1-\lambda)\log(\frac{\phi_{kw}}{p_w}) \tag{7}$$

其中，λ 决定了词项 w 在共主题网络节点 k 中关乎其提升度的权重。如果 λ=0，那么词项完全由其提升度决定，即由词项在主题中的概率 ϕ_{kw} 占据词项在整个文档中的概率 p_w 的比率决定。如果 λ=1，则词项排序完全由主题中词项概率 ϕ_{kw} 决定，文献[9]中建议取 λ=0.6。

3　共主题网络在《大学图书馆学报》中的应用

3.1　题录数据获取及预处理

利用中国知网的“中国学术期刊网络出版总库”，检索期刊名称为“大学图书馆学报”，时间为 1989 年–2015 年。为了计算的便利，按照年度对下载的题录文件进行合并，形成 27 个年度文档，每个文档按照年进行编号。每个文档由若干记录构成，该记录包含题目、作者、单位、关键词、摘要等。为了分析从 1989 年到 2015 年期刊关注的核心研究主题，选取题录文献的标题、关键词、摘要进行研究。步骤如下：

(1) 对 27 个文件做循环读取；

(2) 通过 Rjieba 分词包对标题、关键词、摘要做分词处理，利用正则表达式清理不相干字符(连接符、空值、数字等)；

(3) 使用 tm 包构建文档–词频矩阵，并通过 TF-IDF 优化特征词。后续研究是基于这个词频矩阵开展的。

3.2　共主题网络分析

为了构建主题网络，LDA 必须给出一个主题的最优数目。一般主题数目都是经验给定的，经验给定的问题在于设定的主题数目少了则不能全面表示文档，设定的主题数目多了则主题重复。因此在求证主题数目时大多使用自动发现主题数目的方法。目前有几种自动发现文档中最优主题的方法，包括贝叶斯统计中的标准方法[3]、KL 距离法[32]、余弦相似度法[33]、JSD 法[34]。由于贝叶斯标准统计方法的简洁和计算的效率，已经被大量的研究所使用，因此本实验采用该方法。计算方法见公式(8)和公式(9)，结果如图 3 所示。

$$P(w|z)=\left(\frac{\Gamma(V\beta)}{\Gamma(\beta)^V}\right)^K\prod_{k=1}^{K}\frac{\prod_V\Gamma(n_k^{(w)}+\beta)}{\Gamma(n_k^{(\cdot)}+V\beta)} \tag{8}$$

其中，$n_k^{(w)}$ 是在随机主题 z 中被分配到第 k 个主题的单词 w 的频次。Γ(.) 是标准的 Gamma 函数，$n_k^{(\cdot)}$ 是分配给主题 k 的所有词数。P(w|T) 可以近似为一系列 P(w|z) 的调和平均数，计算公式为：

$$P(w|T)=\frac{1}{M}\sum_{m=1}^{M}P_m(w|z) \tag{9}$$

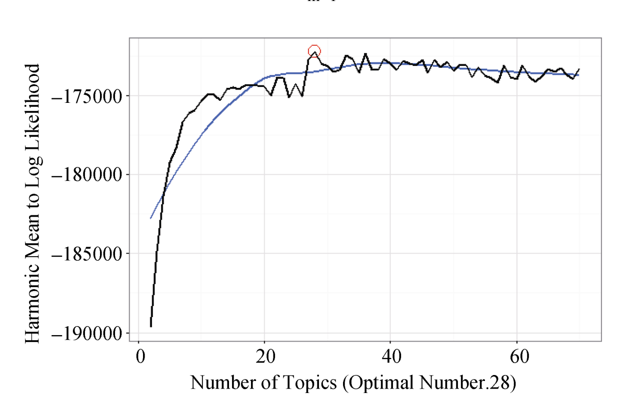


图 3　调和平均数求解最优主题数

选取 k=28 为最大主题数量，建立 27 行 28 列的“文档–主题”矩阵(D×T)，这个矩阵在 LDA 中被命名为 θ，其中 D 为文档，T 为主题，矩阵内容为 T 在 D 中的词项概率分布。按照 LDA 的训练要求，每个文档 d 都是由 k 个主题构成的，但是根据 k 的概率分布情况，总是有若干个重要的主题是代表这个文档的，因此需要选取代表文档 d 的主题，并对它们进行分析。选择规则

是选取文档 d 中大于平均概率的主题作为代表该文档的主题来构建文档–主题二分图。文档主题选取结果如表 1 所示:

表 1　选取的文档主题(部分)

文档(D)	主题(T)
1989	T4 , T23 , T26, T27
2000	T12 , T23 , T26, T27
…	…
2014	T13, T14
2015	T13, T28

对表 1 中数据进行二分图构建，得到文档–主题二分图，如图 4 所示:

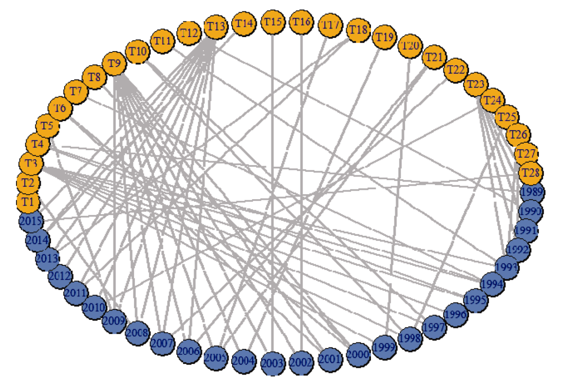


图 4　文档–主题二分图

为更好地说明主题网络中节点的聚类情况，利用 GN 算法对主题网络进行社区分割，通过模块度 Q 判断不同阈值 μ 下产生的社区分割是不是最优，在迭代运算的倒数第 2 次时模块度达到最高的 0.4904142，社区因此被自动切分为两个社区，如图 5 所示:

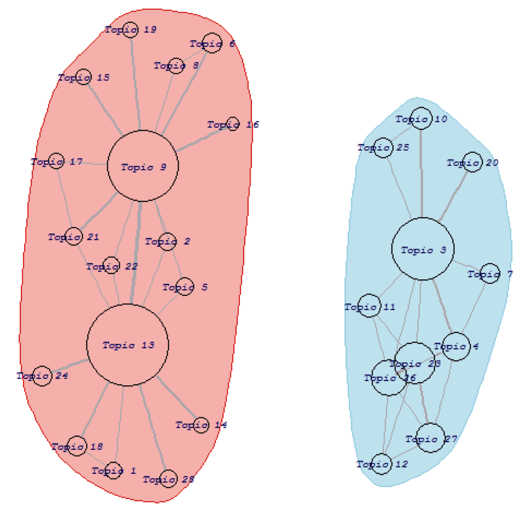


图 5　共主题网络

共主题网络中重要节点为 T13、T3、T9、T23。共主题网络节点之间的权重越大，说明它们在揭示文档时共同出现的机会越多，联系越紧密。节点的面积越大，说明在文档解释中重要性越大。例如 T13 和 T9 的边权较大，说明它们在文档中共同出现的机会很大，而且它们在网络中的节点面积也很大，说明它们是代表文档的重要主题。为了更加清晰地揭示主题重要程度，对 T13、T3、T9、T23 中的前 30 个主题词项云进行观察，如图 6 所示:



图 6　重点主题节点的主词项分布

可以看出，数字、服务质量、图书馆自动化等逐渐成为《大学图书馆学报》期刊关注的重点。

3.3　共主题网络与基于 JSD 的 K-means 比较

Kim 等比较了余弦相似、Jaccard 系数、Kendall τ 系数、DCG 指标、KL 距离、JSD 等测度指标，结果显示 JSD 在主题距离测度上表现最好[35]。为了显示共主题网络在主题关系以及主题重要节点探测方面的能力，将其与表现最好的基于 JSD 的 K 均值聚类进行比较，观察共主题网络的可视化与基于 JSD 的多维标度展示的聚类可视化情况。JSD 是一种基于 KL 距离的度量方法，改善了 KL 距离的不对称问题，成为概率主题的常用测度方法。任意两个主题 T_p 与 T_q 的 JSD 测度公式如下:

$$JSD(\phi_p \| \phi_q) = \frac{1}{2}\left(\sum_{x \in X}\left(\phi_p(x)\log\frac{2\times\phi_p(x)}{\phi_p(x)+\phi_q(x)}\right) + \sum_{x \in X}\left(\phi_q(x)\log\frac{2\times\phi_q(x)}{\phi_p(x)+\phi_q(x)}\right)\right) \quad (10)$$

为了观察 JSD 测度结果和共主题网络测度结果是否一致，分别选择最优主题数 28 和主观选择的主题数 30、20 这三个数目进行测试。对它们进行共主题网络分析，其中社区分割的 Q 值融合如图 7 所示：

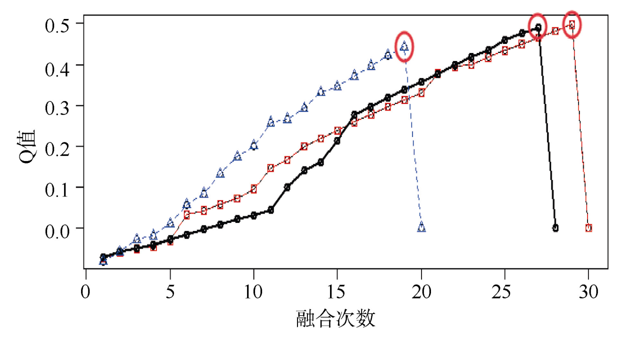


图 7　主题网络 Q 值融合

黑色线为最优主题数 28 的情况，红色线和蓝色线分别为主题数 30 和 20 的情况。它们在 Q 值融合到倒数第 2 次时达到最大值，见红色圆圈处。说明这三个主题数目构成的共主题网络都会聚类成两个社区。最后共主题网络的可视化结果如图 8 上半部分所示。

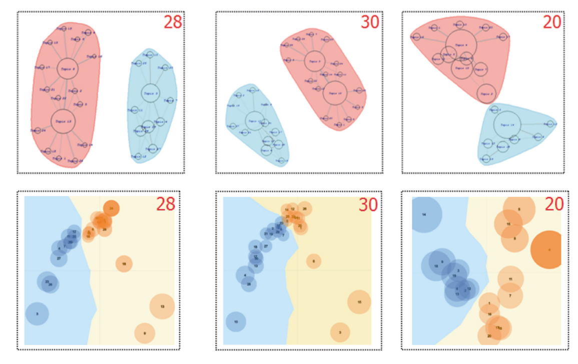


图 8　K-means 聚类对比

使用基于 JSD 的 K-means 聚类同样测度 28、30、20 三种主题数，如果 JSD 测度下的 K-means 聚类和共主题网络下的聚类结果一致，那么就可以判断共主题网络的聚类是可行的。判断基于 JSD 的 K-means 聚类的数目是否与模块度划分的聚类数目一致。基于 JSD 的 K-means 聚类的数目通过轮廓系数(Silhouette Coefficient)方法判断，其值在–1 到+1 之间取值，值越大表示聚类效果越好，最大值对应的聚类数目就是最佳聚类数目[36]。依据这个原理使用多个聚类数目，反复计算每个聚类数目条件下的轮廓系数，当轮廓系数取最大时，其相应的聚类数目是最好的。通过枚举，令聚类数目 k 从 2 到 8，为了避免局部最优解在每个 k 值上重复运行 25 次 K-means，并计算当前 k 的平均轮廓系数，最后选取轮廓系数最大的值对应的 k 作为最终的聚类数目，计算结果如图 9 所示：

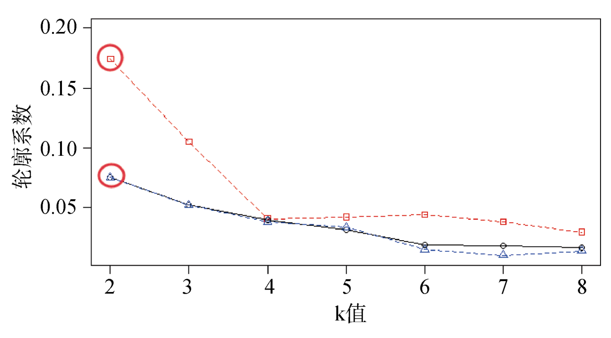


图 9　轮廓系数与 k 的关系

其中，红色、黑色、蓝色分别表示 30、28、20 个主题，三者的最优 k 值都是 2，见红色圆圈标注处。轮廓线划分的聚类主题与模块度划分的主题数目完全相同。选择 k=2 对主题数为 30、28、20 的主题进行聚类(其中主题节点的面积大小与该主题的概率分布值大小相关)，结果如图 8 下半部分所示。分别抽取共主题网络社区分割和 K-means 聚类中内容进行比较，发现基于 JSD 的聚类和共主题网络聚类的结果相似度很高，分别为 28(100%)、20(95%)、30(87%)。比较结果如表 2 所示，其中的数字为主题编号。

表 2　两种方法聚类效果比较

主题数	共主题网络社区划分		基于 JSD 的 K-means 聚类	
	聚类 1	聚类 2	聚类 1	聚类 2
28 个	3, 4, 7, 10, 11, 12, 20, 23, 25, 26, 27	1, 2, 5, 6, 8, 9, 13, 14, 15, 16, 17, 18, 19, 21, 22, 24, 28	3, 4, 7, 10, 11, 12, 20, 23, 25, 26, 27	1, 2, 5, 6, 8, 9, 13, 14, 15, 16, 17, 18, 19, 21, 22, 24, 28
30 个	1, 3, 5, 6, 7, 8, 11, 12, 14, 15, 20, 22, 23, 24, 25, 26, 29	2, 9, 16, 19, 4, 10, 13, 28, 30, 17, 18, 21, 27	1, 2, 3, 5, 6, 7, 8, 9, 11, 12, 14, 15, 16, 19, 20, 22, 23, 24, 25, 26, 29	4, 10, 13, 28, 30, 17, 18, 27, 21
20 个	1, 2, 4, 5, 7, 8, 10, 11, 15, 16, 17, 20	3, 6, 9, 12, 13, 14, 18, 19	1, 4, 5, 7, 8, 10, 11, 15, 16, 17, 20	2, 3, 6, 9, 12, 13, 14, 18, 19

从聚类结果来看，最优主题数28产生的共主题网络社区划分的结果与基于JSD的K-means聚类产生的结果完全一样，原因是最优主题数中的主题不重复，而主观选定的30个和20个主题两种方法在聚类时略有一点偏差，但是大部分是一致的。例如主题数20的主题在共主题网络聚类1中主题2是不一致的，但是通过观察其构成的词项，发现它与和它聚在一起的其他主题讨论的话题非常相似，都是讨论图书馆数字化方面的问题。另外共主题网络由于建立了节点之间的边权关系，能够说明哪些主题节点是共同出现在文档之中来解释文档的，而基于JSD主题聚类的方法只能测度主题是否相似，却无法解释主题之间的这种关系。而且因为有介数中心性测度方法，还能够探测出哪些主题是被大多数文档共同使用的，这是JSD聚类做不到的。

4 结果与讨论

本文通过共主题网络方法处理科技文献，实现了以下目标：

(1) 建立主题之间的网络关系，从而解决了传统LDA生成的主题缺乏联系的问题；

(2) 优化主题词项选择，并遴选出与主题相关的词项，并以词云的形式可视化展现。

共主题网络分析与其他基于主成分的主题关系探查的不同在于照顾到了高维数据的需要，也不存在主题之间距离测度到底选择什么方法的问题，能够探查到哪些主题之间联系紧密且共同出现在对文档的解释中。

在研究过程中存在的不足有：

(1) 由于一篇期刊文献不仅包含文本信息，还包括作者信息，如何结合主题和作者信息分析主题的演变情况？尽管目前AT模型[5]进行了相关研究，但其考虑的是单个作者研究几个主题的问题。如何考虑科学家合作网络研究的主题情况，以便于找到这些知识共同体是后续研究需要探讨的；

(2) 目前科技文献内在结构分析常用的手段是共词网络[37]，共主题网络和共词网络的原理有哪些不同，它们在科技文献分析方面的联系与区别是什么，这些问题也是后续工作要探讨的。

参考文献：

[1] 唐果媛，张薇．国内外共词分析法研究的发展与分析[J]. 图书情报工作，2014，58(22)：138-145. (Tang Guoyuan, Zhang Wei. Development and Analysis of Co-word Analysis Method at Home and Abroad [J]. Library and Information Service, 2014, 58(22): 138-145.)

[2] Blei D M, Ng A Y, Jordan M I, et al. Latent Dirichlet Allocation [J]. Journal of Machine Learning Research, 2003, 3: 993-1022.

[3] Griffiths T L, Steyvers M. Finding Scientific Topics [J]. Proceedings of the National Academy of Sciences of the United States of America, 2004, 101(1): 5228-5235.

[4] Sugimoto C R, Li D, Russell T G, et al. The Shifting Sands of Disciplinary Development: Analyzing North American Library and Information Science Dissertations Using Latent Dirichlet Allocation [J]. Journal of the American Society for Information Science and Technology, 2011, 62(1): 85-204.

[5] Rosen-Zvi M, Griffths T, Steyvers M, et al. The Author-topic Model for Authors and Documents [C]. In: Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence. 2004.

[6] Wang X, McCallum A. Topics Over Time: A Non-Markov Continuous-Time Model of Topical Trends [C]. In: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2006: 424-433.

[7] Blei D M, Lafferty J D. A Correlated Topic Model of Science [J]. The Annals of Applied Statistics, 2007, 1(1): 17-35.

[8] Mimno D. Computational Historiography: Data Mining in a Century of Classics Journals [J]. Journal on Computing and Cultural Heritage, 2012, 5 (1): 1-19.

[9] Sievert C, Shirley K E. LDAvis: A Method for Visualizing and Interpreting Topics [C]. In: Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces. 2014.

[10] Zhang H, Qiu B, Giles C L, et al. An LDA-based Community Structure Discovery Approach for Large-scale Social Networks [C]. In: Proceedings of the 2007 IEEE International Conference on Intelligence and Security Informatics. 2007.

[11] Wang X, Zhang K, Jin X, et al. Mining Common Topics from Multiple Asynchronous Text Streams[C]. In: Proceedings of the 2nd ACM International Conference on Web Search and Data Mining. 2009.

[12] Newman D, Asuncion A, Smyth P, et al. Distributed Algorithms for Topic Models [J]. Journal of Machine Learning Research, 2009, 10(12): 1801-1828.

[13] Gretarsson B, O'Donovan J, Bostandjiev S, et al. TopicNets:

Visual Analysis of Large Text Corpora with Topic Modeling [J]. Transactions on Intelligent Systems & Technology, 2012, 3(2): 565-582.

[14] He Q, Chen B, Pei J, et al. Detecting Topic Evolution in Scientific Literature: How Can Citations Help? [C]. In: Proceedings of the 18th ACM Conference on Information and Knowledge Management. 2009.

[15] Cha Y, Cho J. Social-network Analysis Using Topic Models [C]. In: Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, 2012.

[16] Li D, He B, Ding Y, et al. Community-based Topic Modeling for Social Tagging [C]. In: Proceedings of the 19th ACM International Conference on Information and Knowledge Management. 2010.

[17] Chuang J, Ramage D, Manning C D, et al. Interpretation and Trust: Designing Model-Driven Visualizations for Text Analysis [C]. In: Proceedings of the 2012 SIGCHI Conference on Human Factors in Computing Systems. 2012: 443-452.

[18] Hall D, Jurafsky D, Manning C D. Studying the History of Ideas Using Topic Models [C]. In: Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing. 2008.

[19] Chang J, Boyd-Graber J, Wang C, et al. Reading Tea Leaves: How Humans Interpret Topic Models [R]. Advances in Neural Information Processing Systems 22 (NIPS 2009).

[20] Mimno D, Wallach H M, Talley M, et al. Optimizing Semantic Coherence in Topic Models [C]. In: Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing. 2011.

[21] Latapy M, Magnien C, Del Vecchio N. Basic Notions for the Analysis of Large Two-mode Networks [J]. Social Networks, 2008, 30(1): 31-48.

[22] Newman M E J. Scientific Collaboration Networks. I. Network Construction and Fundamental Results [J]. Physical Review E, 2001, 64(1): 016131.

[23] Zhou T, Ren J, Medo M, et al. Bipartite Network Projection and Personal Recommendation [J]. Physical Review E, 2007, 76(4): 046115.

[24] 任晓龙, 吕琳媛. 网络重要节点排序方法综述[J]. 科学通报, 2014, 59(13): 1175-1197. (Ren Xiaolong, Lv Linyuan. Review of Ranking Nodes in Complex Networks [J]. Chinese Science Bulletin, 2014, 59(13): 1175-1197.)

[25] Newman M E J. Fast Algorithm for Detecting Community Structure in Networks [J]. Physical Review E, 2004, 69(6): 066133.

[26] Girvan M, Newman M. Community Structure in Social and Biological Networks [J]. Proceedings of the National Academy of Sciences, 2002, 99(12): 7821-7826.

[27] Clauset A, Newman M E J, Moore C. Finding Community Structure in Very Large Networks [J]. Phyisical Review E, 2004, 70(6): 066111.

[28] Newman M E J. Modularity and Community Structure in Networks [OL]. ArXiv: physics/0602124v1.

[29] Brandes U, Delling D, Gaertler M, et al. Maximizing Modularity is Hard [OL]. arXiv: Physics/0608255.

[30] Taddy M A. On Estimation and Selection for Topic Models [C]. In: Proceedings of the 18th International Conference on Artificial Intelligence and Statistics.2015.

[31] Bischof J M, Airoldi E M. Summarizing Topical Content with Word Frequency and Exclusivity [C]. In: Proceedings of the 29th International Conference on Machine Learning. Omnipress. 2012.

[32] Arun R, Suresh V, Madhavan V C E, et al. On Finding the Natural Number of Topics with Latent Dirichlet Allocation: Some Observations [A]. // Advances in Knowledge Discovery and Data Mining [M]. Springer Berlin Heidelberg, 2010: 391-402.

[33] Cao J, Xia T, Li J, et al. A Density-based Method for Adaptive LDA Model Selection [J]. Neurocomputing, 2008, 72(7-9): 1775-1781.

[34] Deveaud R, SanJuan E, Bellot P. Accurate and Effective Latent Concept Modeling for Ad Hoc Information Retrieval [J]. Document Numérique, 2014, 17(1): 61-84.

[35] Kim D, Oh A. Topic Chains for Understanding a News Corpus [C]. In: Proceedings of the 12th International Conference on Computational Linguistics and Intelligent Text Processing.2011.

[36] 朱连江, 马炳先, 赵学泉. 基于轮廓系数的聚类有效性分析[J]. 计算机应用, 2010, 32(S2): 139-141. (Zhu Lianjiang, Ma Bingxian, Zhao Xuequan. Clusting Validity Analysis Based on Silhouette Coefficient [J]. Journal of Computer Application, 2010, 32(S2): 139-141.)

[37] 王晓光. 科学知识网络的形成与演化(I): 共词网络方法的提出[J]. 情报学报, 2009, 28(4): 599-605. (Wang Xiaoguang. Formation and Evolution of Science Knowledge Network (I): A New Research Method Based on Co-word Network[J]. Journal of the China Society for Scientific and Technical Information, 2009, 28(4): 599-605.)

利益冲突声明:

作者声明不存在利益冲突关系。

支撑数据:

支撑数据[1-3]由作者自存储, E-mail: niutyut@126.com; 支撑数据[4-7]见期刊网络版 http://www.infotech.ac.cn。

[1] 钮亮. data_preprocessing.R. 《大学图书馆学报》题录文件摘要、标题、关键词的分词处理、文档词频矩阵构建, 最优主题数的确定.

[2] 钮亮. CoTopic_network.R. 三种不同主题数(28, 30, 20)下的文档–主题二分图构建, 加权共主题网络生成, 共主题网络重点主题测度, 模块度计算, 共主题网络社区分割.

[3] 钮亮. JSD_K-means.R. 三种不同主题数(28, 30, 20)下基于JSD的K均值轮廓线系数计算以及主题聚类可视化.

[4] 钮亮. JAL.rar. 知网下载的1989-2015年间《大学图书馆学报》题录文献.

[5] 钮亮. ldamodel.rar. 三种不同主题数(28, 30, 20)下的主题建模数据和文档词频数据, 可供共主题网络和K均值主题聚类使用.

[6] 钮亮. modularity.rar. 三种不同主题数(28, 30, 20)下的模块度数据, 判断最佳社区分割数目.

[7] 钮亮. silhouette-coefficient.rar. 三种不同主题数(28, 30, 20)下的聚类轮廓线数据, 判断最佳聚类数目.

收稿日期: 2016-03-09
收修改稿日期: 2016-05-09

New Research and Application with Co-topics Network

Niu Liang
(School of Economics & Management, China Jiliang University, Hangzhou 310018, China)

Abstract: [**Objective**] This paper builds a co-topics network to analyze the relationship among the topics of research articles and then optimize terms representing these topics. [**Methods**] First, we transformed the "document-topics" bipartite Graph to co-topics networks in accordance with weighted projection rules. Second, we identified the key topics with the combination of betweenness centrality and topic probability. Third, we divided the co-topics network community with the GN algorithm. Finally we optimized topic terms with relevance method. [**Results**] We compared the co-topics networks and the K-means based on JSD by testing optimal topic number (28) and random subjective topic numbers(20, 30). Their clustering numbers were the same and the consistent degree of clustering content reached 100%, 95% and 87%. [**Limitations**] We did not include other community partition methods with the proposed co-topics networks. [**Conclusions**] The co-topics network meets the demands of high-dimensional data and identifies the key topics and the closely linked topics of the target documents.

Keywords: Co-Topics network　LDA　Community partition　K-means